



# Transcriptome analysis

## BIOTECH-7005-BIOINF-3000

Zhipeng Qu

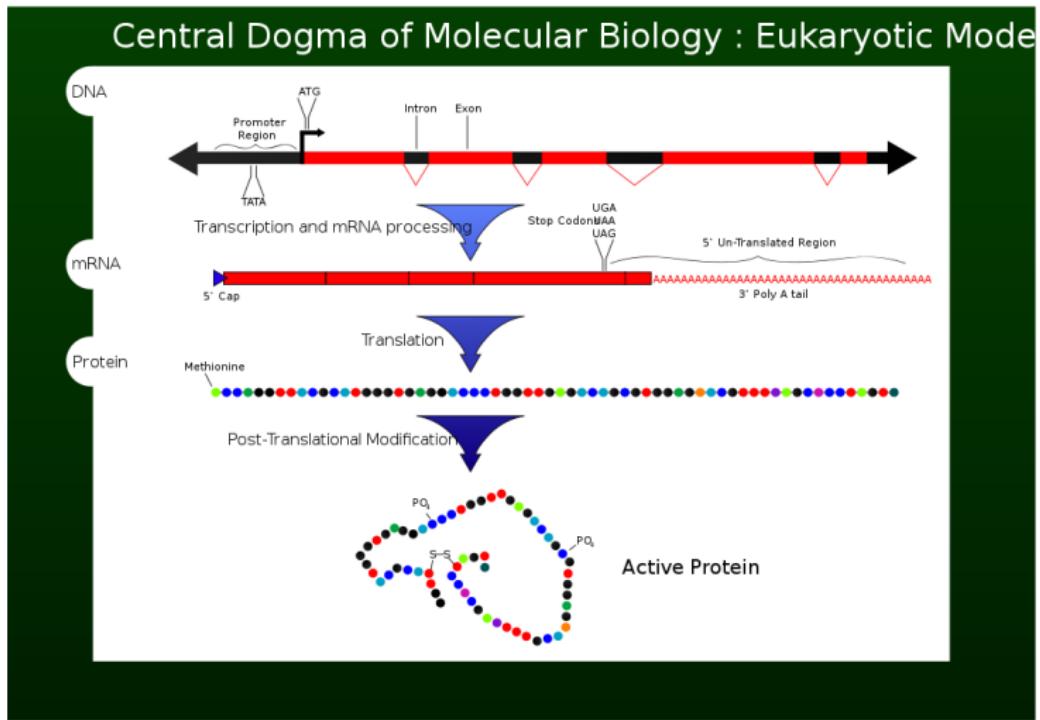
School of Biological Sciences,  
The University of Adelaide

October 17<sup>th</sup>, 2022

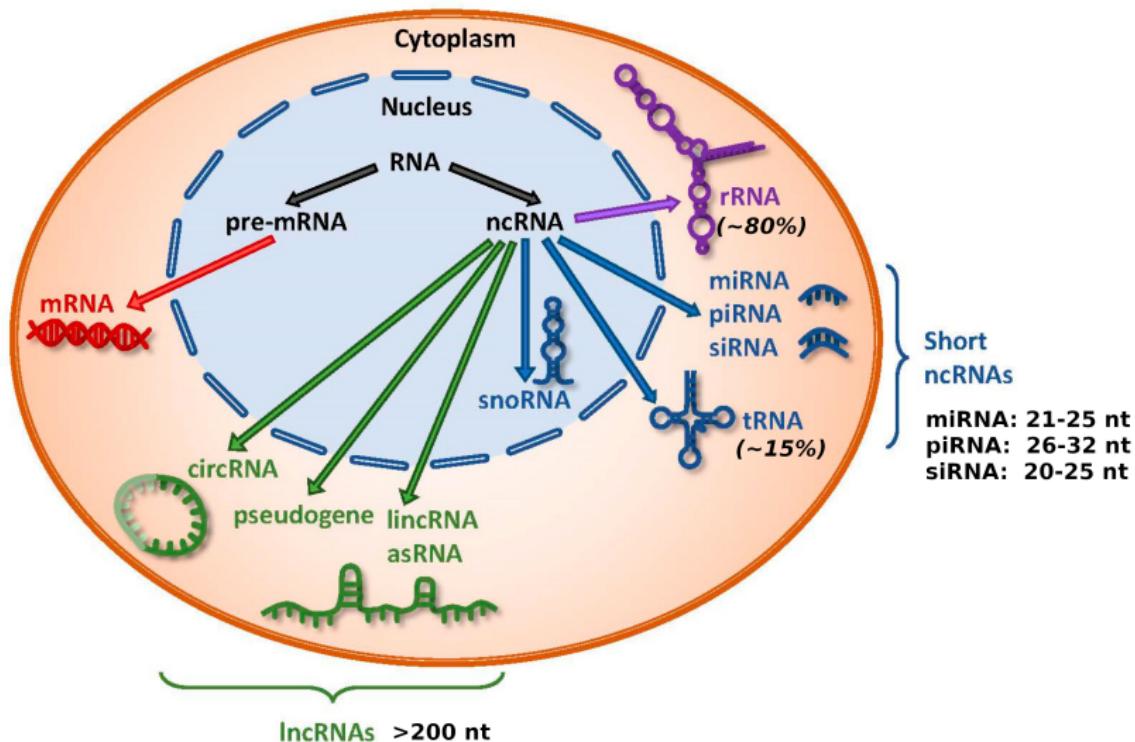
# Outline

- Introduction to transcriptome and transcriptome analysis
- Transcriptome assembly
  - Genome assembly vs transcriptome assembly
  - De novo transcriptome assembly
  - Genome guided transcriptome assembly
  - Assess assembly quality using BUSCO
- Other transcriptome analysis

# Central dogma of biology



# Expression of different RNAs in Eukaryotic cells

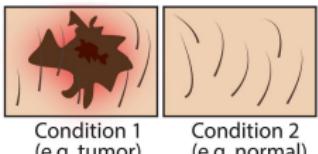


“Transcriptome Analysis is the study of the transcriptome, of the complete set of RNA transcripts that are produced by the genome, under specific circumstances or in a specific cell, using high-throughput methods. ”

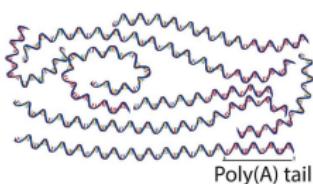
# Transcriptome analysis using RNA-Seq

## Part 1, Library preparation

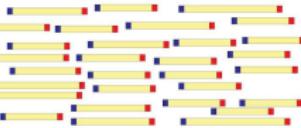
Samples of interest



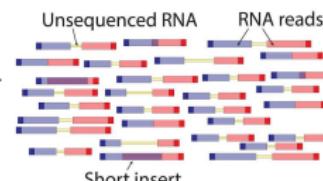
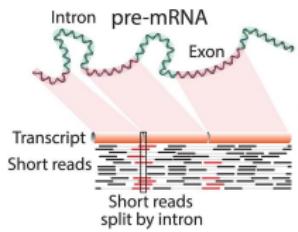
Isolate RNAs



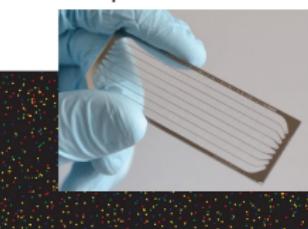
Generate cDNA, fragment, size select, add linkers



## Map to genome, transcriptome, and predicted exon junctions



## Sequence ends



100s of millions of paired reads  
10s of billions bases of sequence

## Downstream analysis

## Part 3, Bioinformatics analysis and Downstream analysis

## Part 2, Next generation sequencing

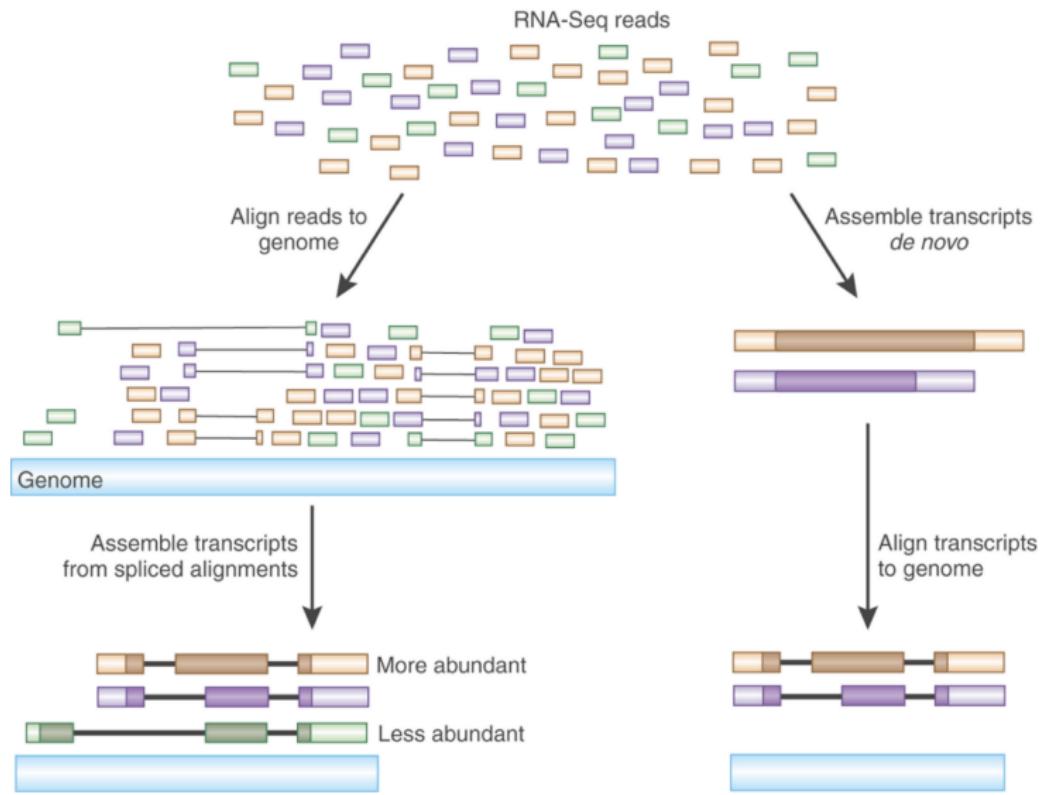
## Different types of transcriptome analysis

- transcriptome assembly
- differential gene expression analysis
- alternative splicing
- non-coding RNAs (ncRNAs)
- other analysis

# Different types of transcriptome analysis

- **transcriptome assembly**
- differential gene expression analysis
- alternative splicing
- non-coding RNAs (ncRNAs)
- other analysis

# Transcriptome assembly using RNA-Seq



## Transcriptome assembly vs genome assembly

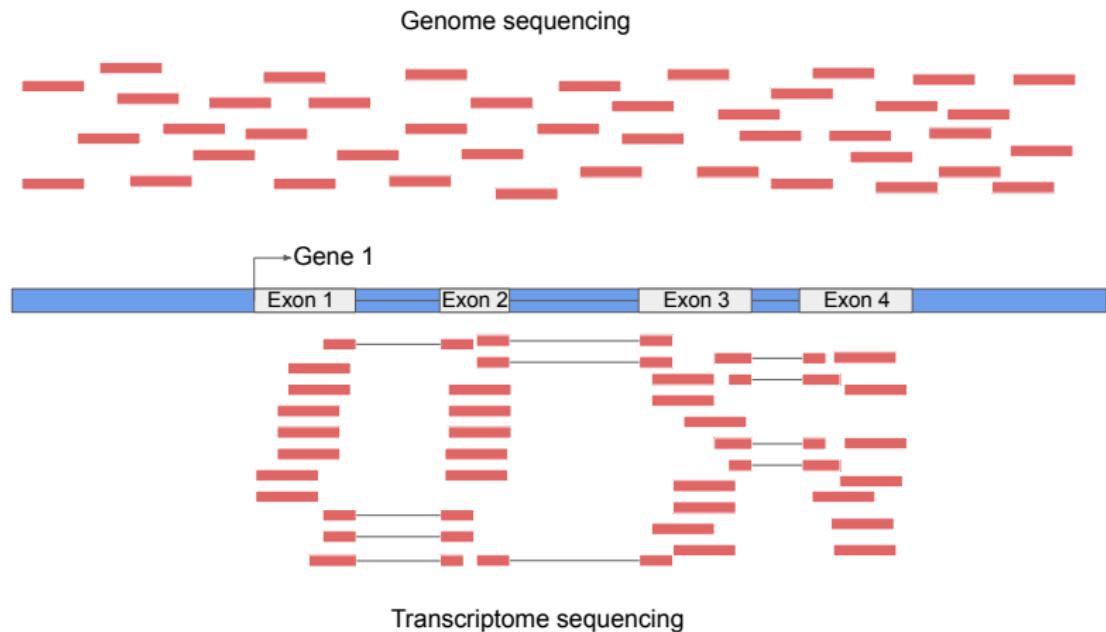
### Genome assembly

- Uniform coverage of genome

### Transcriptome assembly

- Transcribed regions only

# Coverage difference between genome sequencing and transcriptome sequencing



## Transcriptome assembly vs genome assembly

### Genome assembly

- Uniform coverage of genome
- Same genome from one individual

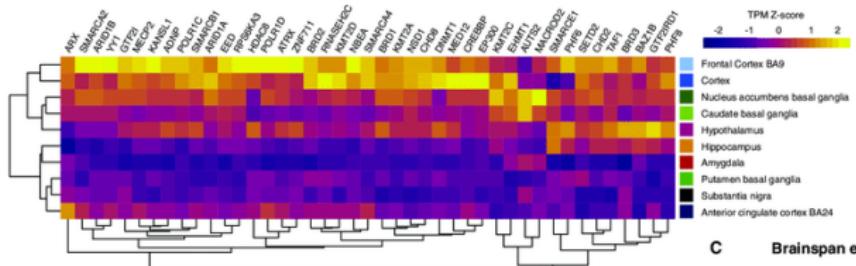
### Transcriptome assembly

- Transcribed regions only
- Temporal- and spatial-specific

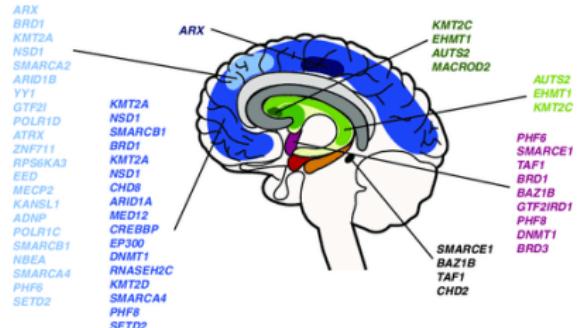
# Temporal- and spatial-specific expression of the transcriptome

**A**

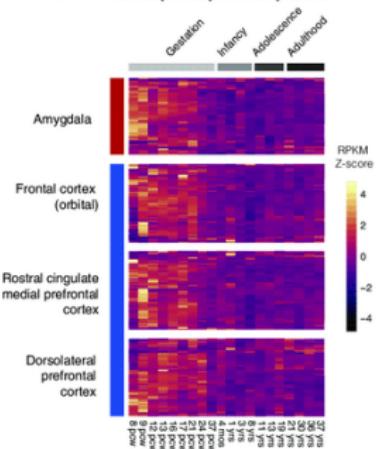
GTEX expression profiles in the adult brain

**B**

Spatial profile of NDD genes in the adult brain

**C**

Brainspan expression profiles



Gabriele, M., Tobon, L. et al., The chromatin basis of neurodevelopmental disorders: Rethinking dysfunction along the molecular and temporal axes. *Progress in Neuro-Psychopharmacology and Biological Psychiatry*, (2017).

## Transcriptome assembly vs genome assembly

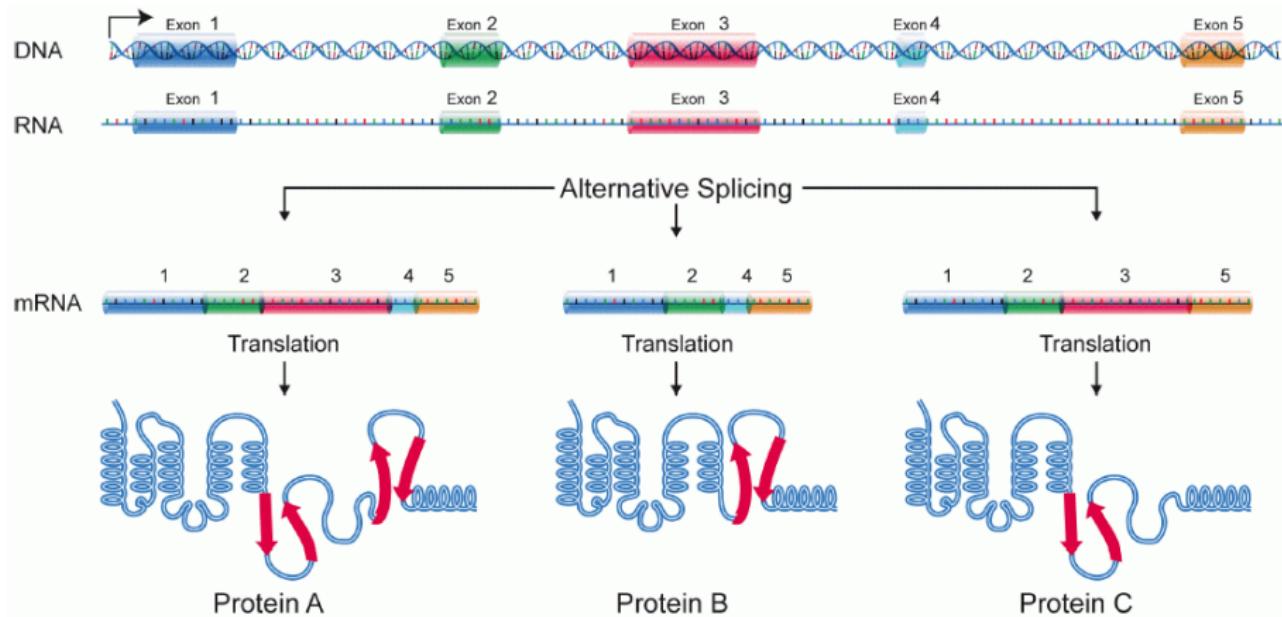
### Genome assembly

- Uniform coverage of genome
- Same genome from one individual
- Single copy per locus

### Transcriptome assembly

- Transcribed regions only
- Temporal- and spatial-specific
- Single/Multiple copy(ies) per locus (Alternative splicing)

# Alternative splicing in the transcriptome



## Transcriptome assembly vs genome assembly

### Genome assembly

- Uniform coverage of genome
- Same genome from one individual
- Single copy per locus
- Assemble small number of large contigs (Mb)

### Transcriptome assembly

- Transcribed regions only
- Temporal- and spatial-specific
- Single/Multiple copy(ies) per locus (Alternative splicing)
- Assemble thousands of shorter transcripts (Kb)

# Transcriptome length distribution

## Transcript distribution

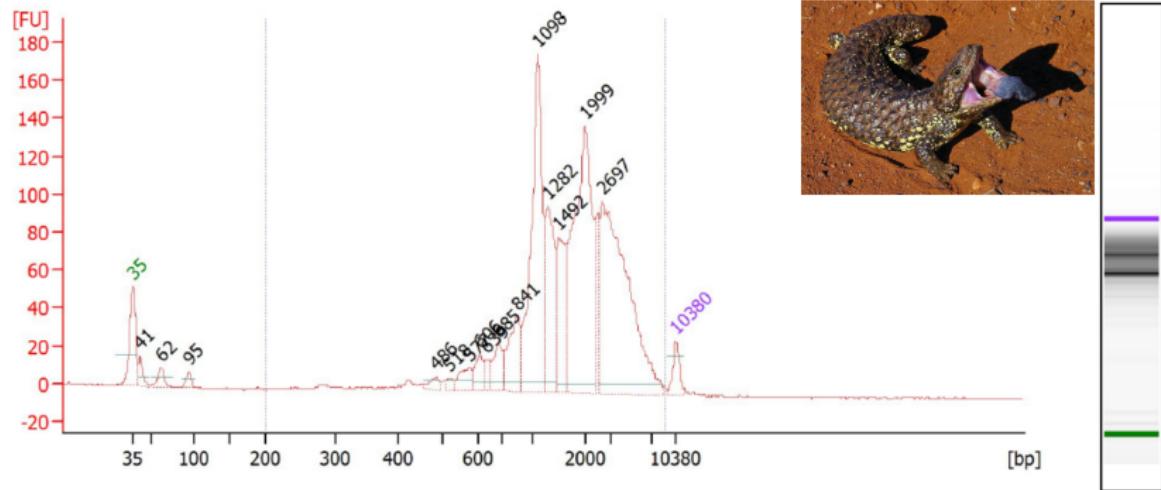


Figure courtesy of Dr Terry Bertozzi

## Transcriptome assembly vs genome assembly

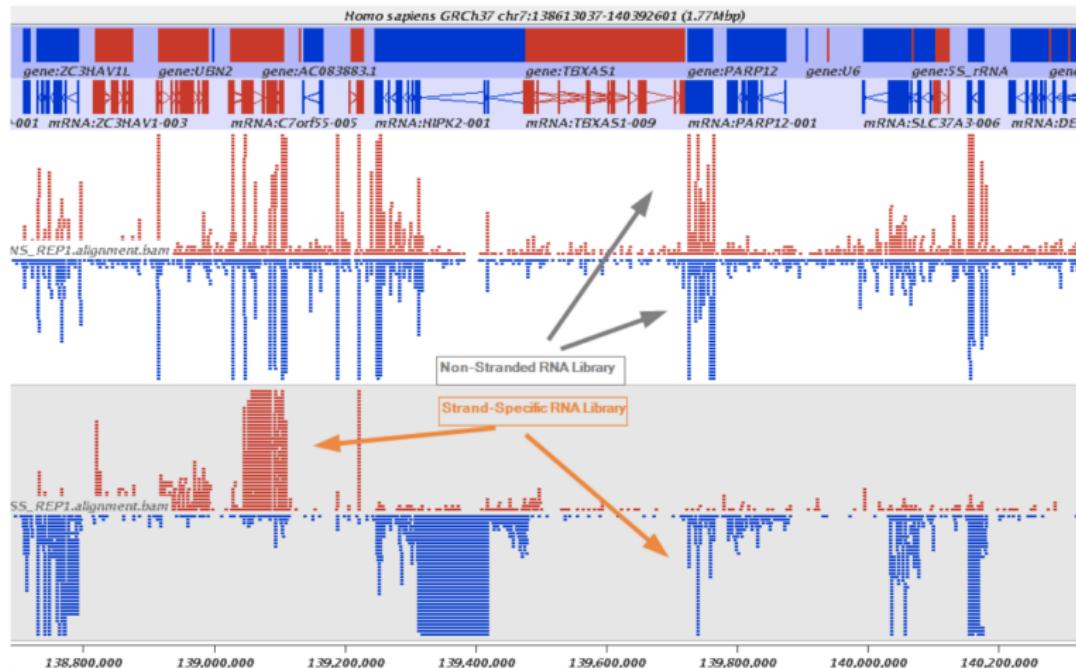
### Genome assembly

- Uniform coverage of genome
- Same genome from one individual
- Single copy per locus
- Assemble small number of large contigs (Mb)
- Double stranded

### Transcriptome assembly

- Transcribed regions only
- Temporal- and spatial-specific
- Single/Multiple copy(ies) per locus (Alternative splicing)
- Assemble thousands of shorter transcripts (Kb)
- Strand specific

# Strand specific expression in transcriptome



<https://training.galaxyproject.org/training-material/topics/transcriptomics/tutorials/rb-rnaseq/tutorial.html>

## *De novo transcriptome assembly*

Why do we need *de novo* transcriptome assembly?

- No available reference genome (non-model organisms)
- Reference genome is not in high quality
- Cheaper than whole genome assembly

## *De novo transcriptome assembly - Trinity*



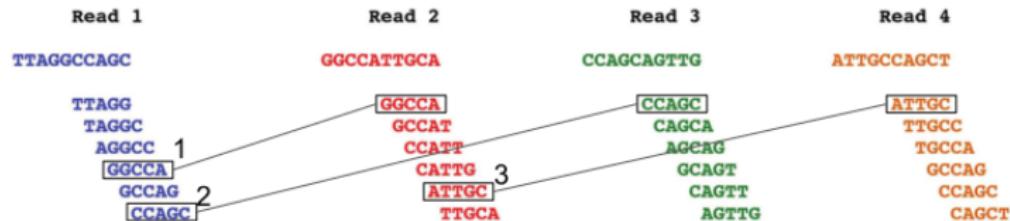
- Developed at the Broad Institute and the Hebrew University of Jerusalem
- *De Bruijn* graphs based
- Three independent software modules: Inchworm, Chrysalis, and Butterfly
- Outputs a FASTA format sequence file

---

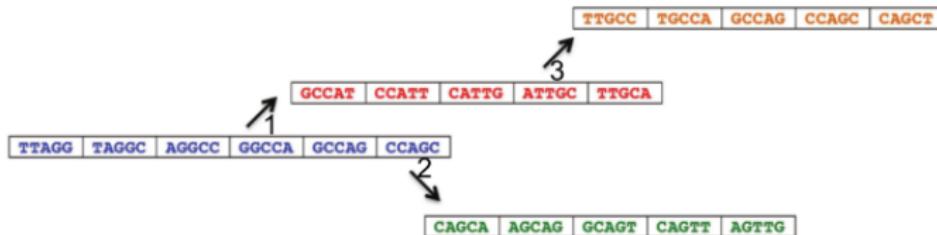
<https://github.com/trinityrnaseq/trinityrnaseq/wiki>

# NGS assembly using K-mer and *De Bruijn* graph

k-mer size: 5



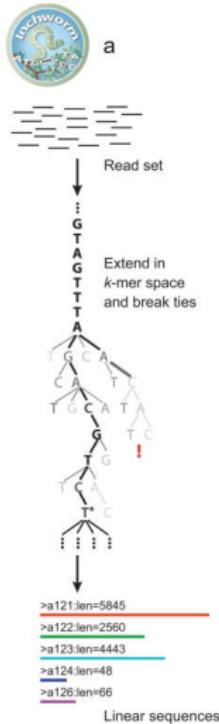
De Bruijn graph



Reconstructed transcripts

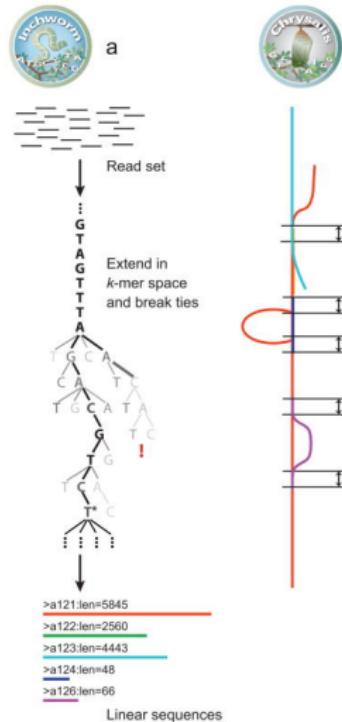
- Isoform 1 TTAGGCCATTGCA
- Isoform 2 TTAGGCCATTGCCAGCCT
- Isoform 3 TTAGGCCAGCAGTTG

## Trinity - step 1



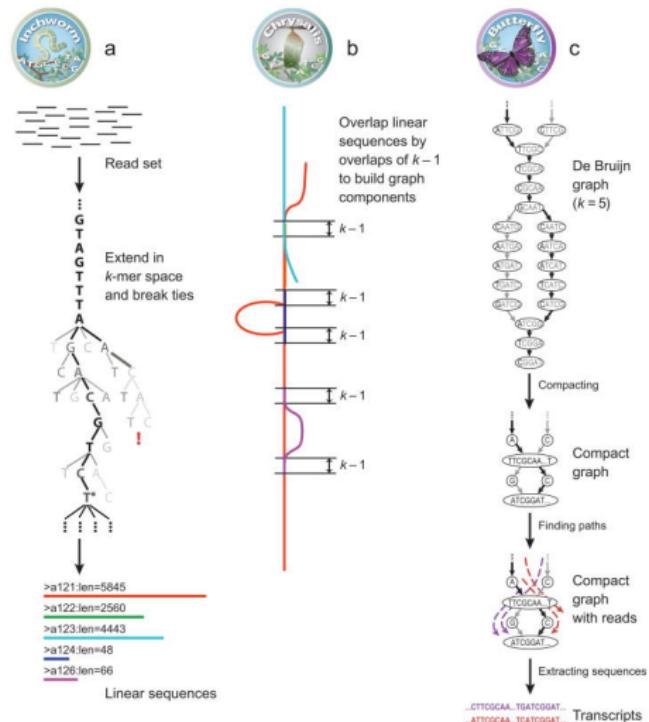
Grabherr M., et al., Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2021 Jul; 29(7): 644-652

## Trinity - step 2



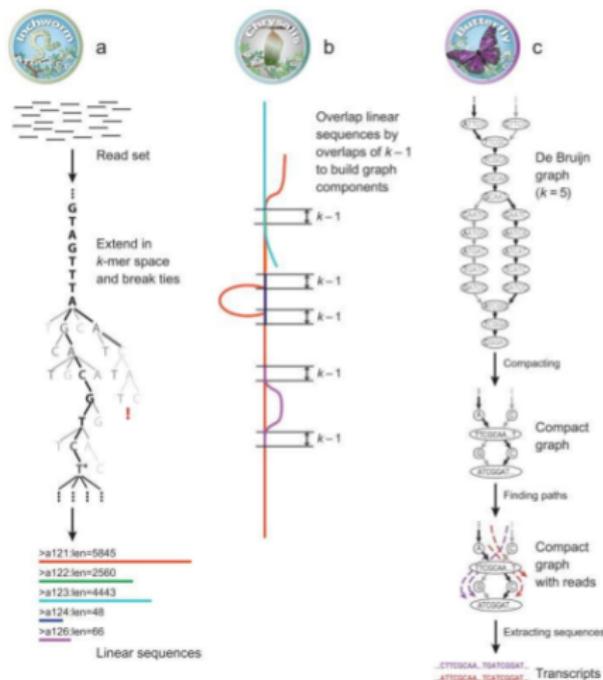
Grabherr M., et al., Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2021 Jul; 29(7): 644-652

## Trinity - step 3



Grabherr M., et al., Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2021 Jul; 29(7): 644-652

# Trinity - Output

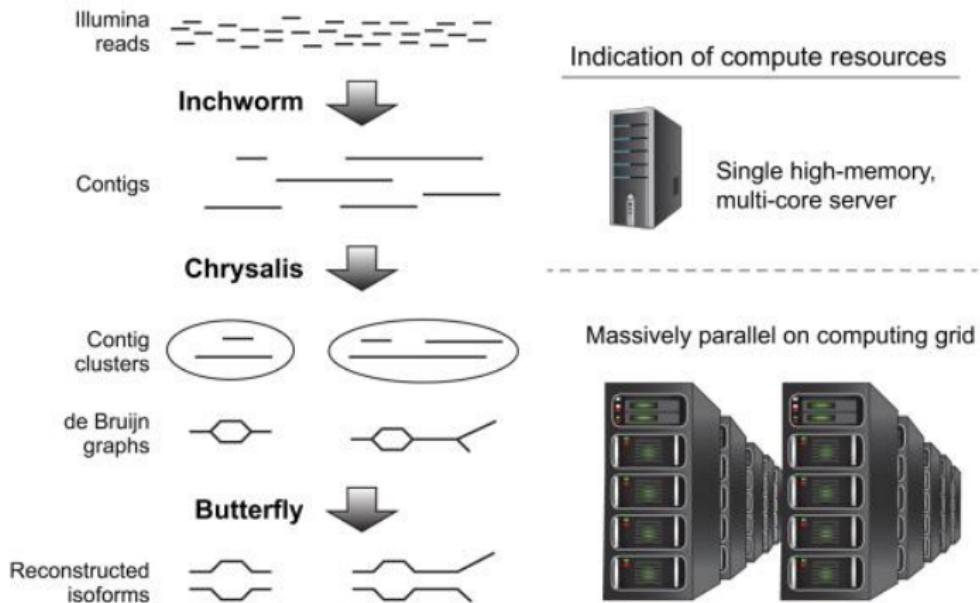


Output: Trinity.fasta

```
>comp0_c0 seq1 len=5528 path=[3647:0-3646 129:3647-3775 1752:3776-5527]
AATTGAATCCCTTTTGATTGAAAAGTTGAATGAAAGACATACAGAT
TGAATGTG_TCCTCTGATACACAGCCTCGCAGGGTCTTCAAGCCGTGG
GCTGCCACGGGTGCTAAGTCAACTGCATTGATGCCGCTTTAACCCCCC
AGGGGACACCTCGGCAGCTGTTGCCCTGAGTA..TTGTGTTCTCAAACAG
TTTACAGCTTGTGATTGCCATTTTATTATTCATTAATCAAGATAATCG
TAAATGGGGGGAGGGCGCCGTCGTAGGGTCTGCACATGGCCCCCGCTCG
CCATGATGACAAGCGCAGAACCTCAGT
>comp0_c0 seq1 len=5399 path=[3647:0-3646 1752:3647-5398]
AATTGAATCCCTTTTGATTGAAAAGTTGAATGAAAGACATACAGAT
TGAATG..TGGTATTGCAAATATAATGCAATTTCGAAACAATTAAATTATG
AAAATATAC..TTGTGTTCTTCAACAGTTTACAGCTTGTGTAATTGCCATT
TTATTATTCATTATCAAGATAATCGTAAATGGGGGGAGGGCGCCGCTGTG
TAGGGTCTGCACATGGCCCCCGCTGCATGATGACAAGCGCAGAACCTCA
GT
```

Grabherr M., et al., Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol. 2021 Jul; 29(7): 644-652

# Trinity - required compute resources

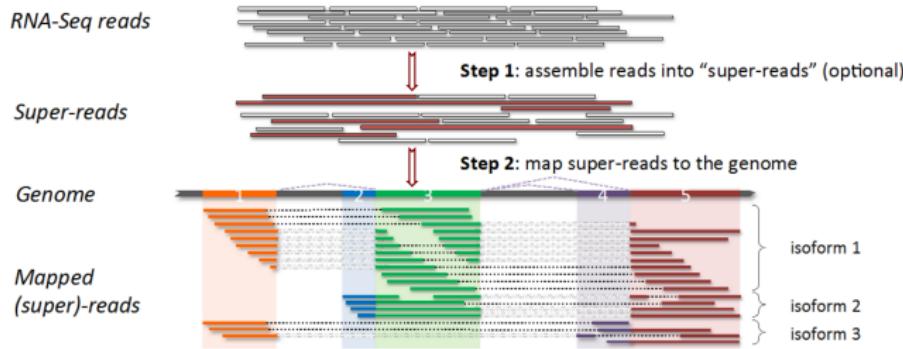


Haas B., et al., De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc. 2013 Aug; 8(8)

## When do we use genome guided transcriptome assembly?

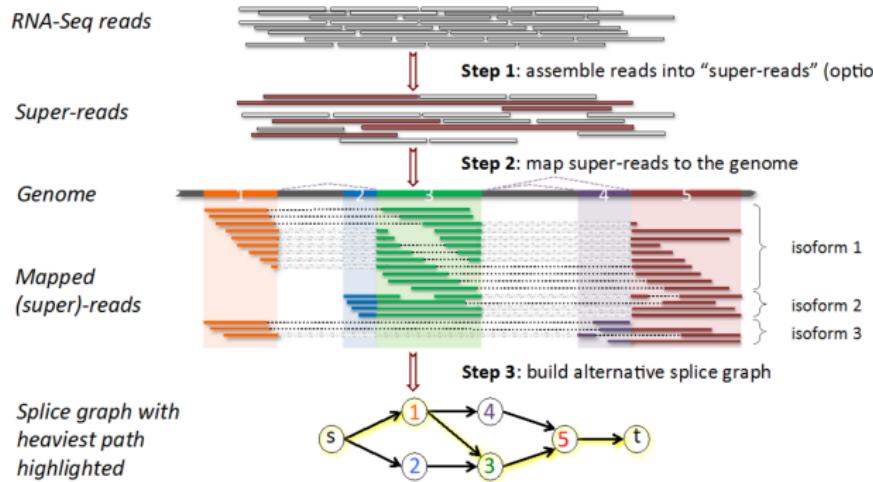
- With an available reference genome
- To improve current gene annotation
- To identify alternative splicing isoforms
- To identify novel transcripts (long ncRNAs)

# StringTie



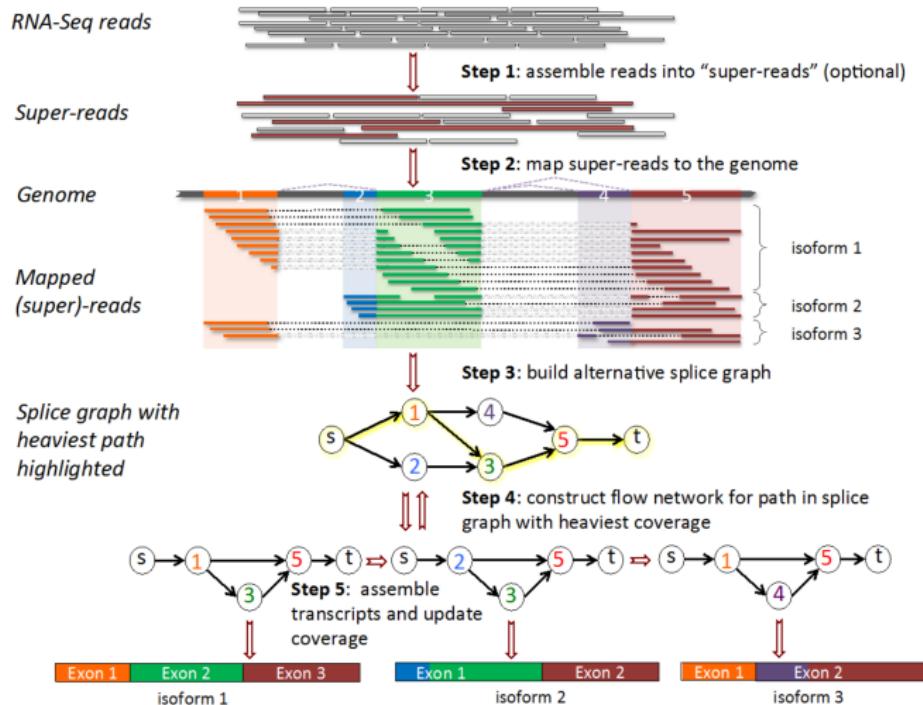
Pertea M., et al., StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015 Mar; 33(3): 290-295

# StringTie



Pertea M., et al., StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015 Mar; 33(3): 290-295

# StringTie



Perteau M., et al., StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat Biotechnol. 2015 Mar; 33(3): 290-295

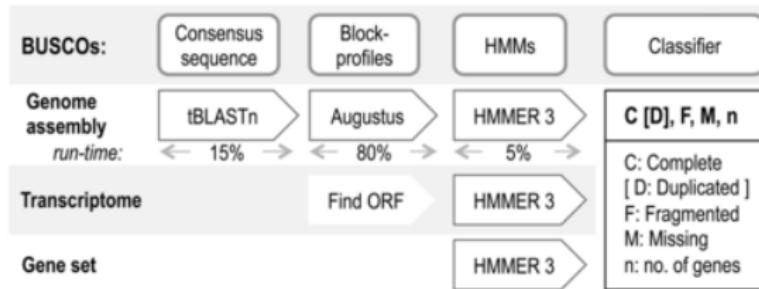
## StringTie

```
seqname source      feature    start    end      score    strand   frame attributes
chrX   StringTie   transcript  281394   303355  1000     +        .       gene_id "ERR188044.1";
chrX   StringTie   exon       281394   281684  1000     +        .       gene_id "ERR188044.1";
...
...
```

- StringTie outputs a Gene Transfer Format (GTF) file that contains details of the transcripts that StringTie assembles from RNA-Seq data. The definition of each column in this GTF file can be found at ENSEMBL.
- StringTie can also calculate coverage for expression analysis.

## Assessing assembly completeness - BUSCO

- Benchmarking Universal Single-Copy Orthologs (BUSCO)
- Can be used on both genome and transcriptome assemblies



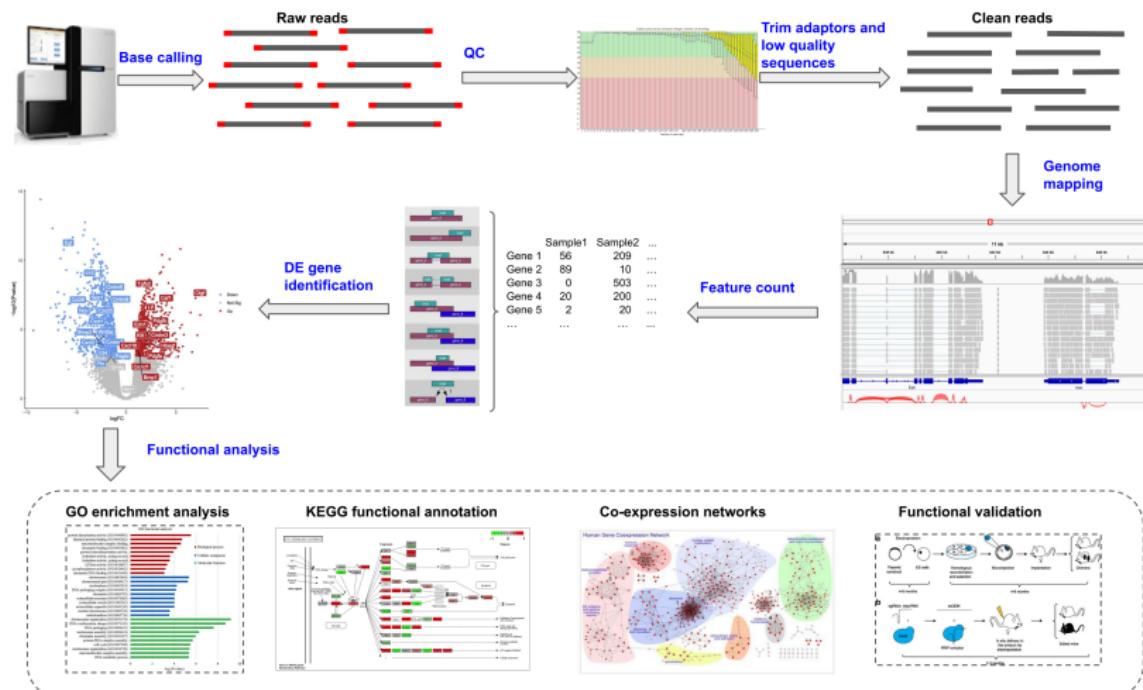
Simao F., et al., BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015 October; 31(19): 3210-3212

## Assessing assembly completeness - BUSCO report

**C:89.0%[S:85.8%,D:3.2%],F:6.9%,M:4.1%,n:3023**

- C: Complete orthologs
  - S: Single copy
  - D: Duplicated copy
- F: Fragmented copy
- M: Missing copy
- n: Total orthologs

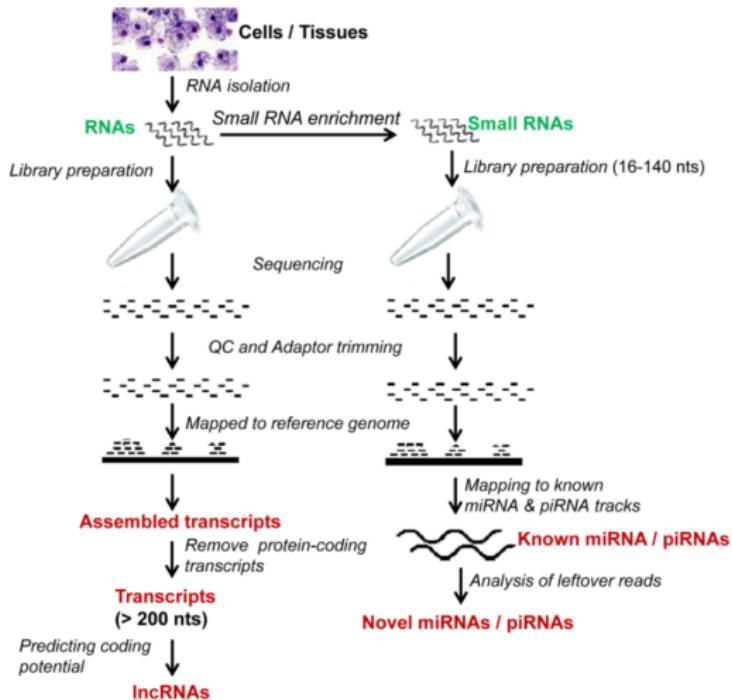
# Other transcriptome analysis - Differential gene expression analysis



# Other transcriptome analysis - Alternative splicing



# Other transcriptome analysis - non-coding RNAs (ncRNAs)



Mallick (2016) Decrypting the Treasures of Regulatory Non-coding RNAs in High-throughput Era. J Data Mining Genomics and Proteomics 7: e124.

# Lecture summary

- Transcriptome analysis is the study of the total RNA readout
- Transcriptome assembly has many challenges compared to genome assembly
- De novo transcriptome assembly - Trinity
- Genome guided transcriptome assembly - StringTie
- Assessing assembly using BUSCO